VOICE OF SISYPHUS: AN IMAGE SONIFICATION MULTIMEDIA INSTALLATION

Ryan McGee, Joshua Dickinson, and George Legrady

Experimental Visualization Lab
Media Arts and Technology
University of California, Santa Barbara
ryan@mat.ucsb.edu, dickinson@mat.ucsb.edu, legrady@arts.ucsb.edu

ABSTRACT

Voice of Sisyphus is a multimedia installation consisting of a projection of a black and white image sonified and spatialized through a 4 channel audio system. The audio-visual composition unfolds as several regions within the image are filtered, subdivided, and repositioned over time. Unlike the spectrograph approach used by most graphical synthesis programs, our synthesis technique is derived from raster scanning of pixel data. We innovate upon previous raster scanning image to sound techniques by adding frequency domain filters, polyphony within a single image, sound spatialization, and complete external control via network messaging. We discuss the custom software used to realize the project as well as the process of composing a multimodal artwork.

1. INTRODUCTION

Voice of Sisyphus relies on an Eisensteinian process of *montage* [1], the assembling of phrases with contrasting visual and tonal qualities as a way to activate change that can be considered as a narrative unfolding. Whereas cinematic montage involves the contrast of discontinuous audio-visual sequences as a way to build complexity in meaning, in this work, the referent photograph that is processed does not ever change, except through the filtering that generates the tonal and visual changes. As the composition evolves but then returns to where it began, the event brings to mind the Greek myth of king Sisyphus who was compelled to ceaselessly roll an immense boulder up a hill, only to watch it roll back down repeatedly. The intent is to have a continuously generated visual and sound composition that will keep the spectator engaged at the perceptual, conceptual, and aesthetic levels even though the referent visual source is always present to some degree.

Voice of Sisyphus was partially inspired by the overlay of image processing techniques in Peter Greenaways 2009 film, *Wedding at Cana*[1], a multimedia installation that digitally parses details of the 1563 painting by the late Renaissance artist Pablo Veronese in a 50 minute video. The filmmaker skillfully uses computer vision techniques to highlight, isolate, and transform visual details to explore the meaning of the visual elements in the original painting.

The project evokes two early digital works by Legrady. *Noise-To-Signal*[2] (1986) is an installation artwork that uses digital processing to explore the potential of image analysis, noise, and Information Theory's definition of noise to signal. *Equivalents II*[3], realized in 1992, is another interactive digital media artwork that implements 2D midpoint fractal synthesis as a way to create organic-looking abstract images whose abstract cloud-like visual forms were defined by textual input provided by viewers. Both artworks integrated synthesis algorithms to generate cultural content through computational creation of images.

Most experiments examining the relationships between sound and image begin with sounds or music that influence the visuals. Chladni's famous 18th century "sound figures" experiment involves visual patterns generated by playing a violin bow against a plate of glass covered in sand[2]. 20th century visual music artists often worked by tediously synchronizing visuals to preexisting music. Though, in some cases, the sounds and visuals were composed together as in *Tarantella* by Mary Ellen Bute. Today, visual artists often use sound as input to produce audio-reactive visualizations of music in real-time.

Less common are technical methodologies requiring images as input to generate sound. However, in 1929 Fritz Winckel conducted an experiment in which he was able to receive and listen to television signals over a radio[2], thus resulting in an early form of image audification. Rudolph Pfenninger's *Tnende Handschrift* (Sounding Handwriting), Oskar Fischinger's *Ornament Sound Experiments*, and Norman McLaren's *Synchromy* utilized a technique of drawing on film soundtracks by hand to synthesize sounds. *Voice of Sisyphus* continues in the tradition of the aforementioned works by using visual information to produce sound.

2. SOFTWARE

Custom software was developed to realize the artist's vision of translating an image into a sonic composition. Although *Voice of Sisyphus* is based on a particular photograph, the software was designed to be used with any image. Once an image file is imported one may select any number of rectangular regions within the image as well as the entire image itself to sonify. Greyscale pixel values within a region are read into an array, filtered, output as a new image, and read as an audio wavetable. The wavetables of multiple regions are summed to produce polyphonic sound. Consideration was taken for real-time manipulation of region locations and sizes during a performance or installation without introducing unwanted audio artifacts.

[1] http://www.factum-arte.com/eng/artistas/greenaway/veronese_cana.asp

[2] http://www.mat.ucsb.edu/g.legrady/glWeb/Projects/noise/noise.html

[3] http://www.mat.ucsb.edu/g.legrady/glWeb/Projects/equivalents/Equi.html

2.1. Related Work

Realization of *Voice of Sisyphus* necessitated the development of custom software for our approach to image sonification. A vast majority of existing image sonification software uses the so-called "time-frequency" approach [3] in which an image acts as the spectrograph for a sound. These systems include Iannis Xenakis' UPIC and popular commercial software such as MetaSynth and Adobe Audition. Their shared approach considers the entire image much like a musical score where the vertical axis directly corresponds to frequency and the horizontal axis to time. Usually the image is drawn, but some software like Audition allows the use of bitmap images and considers color as the intensity of frequencies on the vertical axis. MetaSynth uses the color of drawn images to represent the stereo position of the sound. In any case, all of the aforementioned software reads images left to right at a rate corresponding to the tempo. Reading an entire image left-to-right as a means to image sonification has been termed as *scanning* by Yeo[4].

However, our approach was to focus on different regions within an image over the course of the composition. Yeo has termed this approach *probing* [4]. Thus, unlike scanning, the horizontal axis of the image is not related to time. The composer must move or *probe* different regions of the image to advance the time of the composition. We also sought a more literal translation of images to sound than the typical spectrograph scanning approach. We felt that, although novel in their own right, spectrograph scanning approaches adhere too closely to a traditional musical score. We wanted a departure from the common practice of viewing images as time-frequency planes and sought a technique to listen to variations between different regions of an image. We wanted the resulting composition to unfold like one perceives a photograph in a non-linear fashion– first noticing some region, person, or object and then shifting the focus to other objects within the scene.

To produce a literal translation of image regions to sounds we began by looking at the pixel data itself. One convenient constraint was that the image chosen by the artist for the project was black and white so we did not have to consider color in our sonification approach. We began with a straight-forward audification of the 8-bit greyscale pixel values rescaled to be floating-point audio samples. The pixel values are read via raster scanning, that is line by line, top down into a 1 dimensional array of audio samples. We were aware of similar image sonification work by Yeo and Berger [5], but only became aware of their software interface, Rasterpiece [6], after we completed *Voice of Sisyphus*. Rasterpiece allows for regions of an image to be converted to sound via raster scanning with in-between filtering, a process similar to our own which we describe in later sections of this paper. As we will also detail in later sections, our software adds a more desirable filtering technique, multiple regions within the same image, Open Sound Control[7], removal of unwanted sound artifacts when manipulating regions, and sound spatialization.

2.2. Interface

The interface has both editing and presentation modes. Editing mode displays a panel of sliders for manipulating region parameters and clearly outlines all active regions within the image with colored rectangular boxes. One may create, remove, reposition, or resize regions via the mouse. Presentation mode removes the panel of sliders and region outlines from sight, making the application suitable for an artistic installation or performance to be controlled via Open Sound Control (OSC).

Interactive sonification has been defined as "the discipline of data exploration by interactively manipulating the data's transformation into sound."[8] Our software's ability to drastically change the sound obtained from image regions through interactive manipulation of spectral amplitude thresholds and segmentation of regions into a melody of subsections (both described in section 2.3) could be classified as a form of interactive sonification. The composition process for the resulting artwork described in section 3 involved interactive adjustment of parameters within a given model defined by the composer. While the final artwork is not interactive, the process of its creation could be described as working with a model-based sonification[9], which is interactive by definition.

Figure 1: Software Interface for *Voice of Sisyphus*

2.3. Sound Synthesis from Image Data

Currently, our software only deals with 8-bit greyscale images, and any color or other format images imported to the software will first be converted to 8-bit greyscale. The synthesis algorithm begins with a back-and-forth, top-down raster scanning of the greyscale pixel values, which range from 0 to 255 (black to white respectively). Simply scaling these values to obtain floating-point audio samples in the -1.0 to 1.0 range results in harsh, noisy sounds without much variation between separate regions in most images. These initial noisy results were not at all surprising given that the greyscale variation of an arbitrary image will contain a dense, broad range of frequencies. For instance, given a picture of a landscape, analyzing variations in each pixel value over a region containing thousands of blades of grass would easily produce a noisy spectrum with no clear partials. Of course, images can be specifically produced to contain particular spectra and result in tonal sounds [5], but we were interested in exploring the sounds resulting from different regions of any arbitrary image. In our case, *Voice of Sisyphus* uses an evocative photograph of a formal gala reception, so we might ask "What does a face sound like compared to a window?" Of course, the ability to determine high-level descriptions of image regions such as a "face" or "window" is a problem of feature recognition in computer vision, but we were interested in examining the objective differences in the pixel data of a "face" or "window" rather than what sounds we might normally associate with each of those objects. So, such high-level descriptions were

not necessary. We took a spectral-based approach to analyzing and processing each region's pixel data so that we could filter image regions to produce less noisy sounds with greater distinguishability between regions.

We applied a selection of frequency domain filters to our audification of pixel data by implementing a short-time Fourier transform (STFT) for each region. The STFT is obtained by computing a fast Fourier transform (FFT) of each region at the graphics' frame rate. Each FFT gives us amplitudes and phases for frequencies contained in that region at that time. Manipulation of these amplitudes and phases allows us to control the the spectrum of the image and, therefore, the resulting sound in real-time. Zeroing the amplitudes of frequencies above or below a cutoff produces a low-pass or high-pass filter respectively, while scrambling the phases of an FFT scrambles the pixels in an image without affecting is spectrum. Our key filter was to remove all frequencies below a variable amplitude threshold, leaving only the most prominent partials present and, thus, accentuating tonal differences between regions within a single image. Implementing this threshold denoises the resulting sounds, leaving clear tones that change as the region is moved or resized. The pixel data of regions is continuously updated to show the effect of the filters so the observer is always seeing and hearing the same data. As the sound becomes clearer from the filter's removal frequencies, the image becomes blurry. An interesting conclusion from this process is that most perceptually coherent images sound like noise while perceptually clear, tonal sounds result from very abstract or blurry images. This imposed a challenge for the composer of *Voice of Sisyphus* as he describes in later sections of this paper.

To obtain the final image and sound data after applying filters in the frequency domain we compute an inverse short-time Fourier transform (ISTFT) for each region, which gives the filtered pixel values. These new values are then scaled to the range -1.0 to 1.0 and read as an audio wavetable via scanned synthesis, a technique that can be used to scan arbitrary wavetables of audio data at variable rates using interpolation[10]. A control for the scan rate of these wavetables affects the fundamental pitch of the resulting sounds. However, the perceived pitch also changes as regions are moved and resized, causing new partials appear and disappear from the spectrum.

Before computing the FFT we can also scale the pixel data to effect the brightness of the resulting image and, therefore, amplitude of the sound. A masking effect can also be applied at this point, which acts as a bit reduction to the image and sound by quantizing amplitude values. Overall, it is important to note that the software only manipulates the image data and not the audio data. Since the audio data is continually produced in the same manner (scanning the IFFT results), changes in the sound are always directly produced from changes in the image. Simply put, in *Voice of Sisyphus* we are always seeing and hearing the same data. Figure 3 summarizes the sonic effects of the image filters.

The composition dictates the rapid movement and resizing of specific regions which caused discontinuities in our wavetables, resulting in an unwanted audible popping noise. To account for the resizing of images, all resulting audio wavetables, originally a length equal to the number of pixels in an image region, are upsampled or downsampled to a fixed size before linear interpolation is used to read the table at the desired frequency. Wavetables are then cross-faded with each other at the audio buffer rate to prevent discontinuities from the dynamically changing wavetables resulting from the movement and resizing of regions. If the region's position

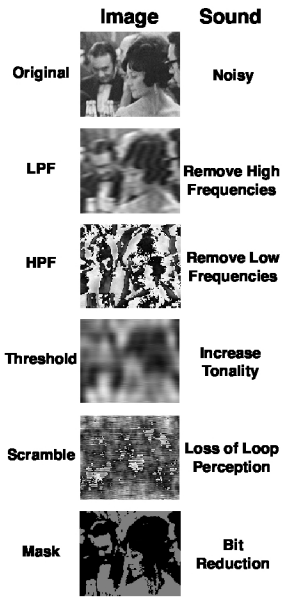

Figure 3: Effects of Image Processing on Sound

and size remain static, then the wavetable is simply looped. Scrambling the phases of the image region during the spectral processing effectively scrambles the time information of our wavetable without altering its spectra and the result is a perceptually continuous rather than looped sound.

Another challenge imposed by the composition was the desire to listen to the entire image at once. Using our STFT technique with N-point FFTs in which N is the number of pixels in the image meant taking over 1 million point FFTs at frame rate for images greater than a megapixel in size. Such computations were not suitable for the desired real-time operation. To solve this problem we added a segmentation mode for large regions which automatically subdivides them into several smaller regions of equal size. The sounds from these regions are then played-back successively left-to-right and top-down. The result of this is quite interesting–the segmentation technique is reminiscent of the step sequencers found in common electronic music hardware. Moving the segmented region produces different melodies from the resulting tones of each subsection of a region. The software's tempo slider controls the rate at which each subsection is played. Applying filters to the regions can also lead to rests in the patterns.

Regions' sounds are spatialized according to their location within the image. If a region is segmented, then the spatialization algorithm updates the position of the sound as each subsection is played, thus adding a spatial component to the aforementioned sequencer. Our method of spatialization is similar to that used in vOICe[11], an augmented reality project for the totally blind– a way to see with sound. In vOICe sounds are spatialized in 1 dimensional stereo according to their pixels' position in the horizontal image plane. *Voice of Sisyphus* uses a 2 dimensional sound plane to spatialize sounds based on to their pixels' horizontal and vertical position in the image. The installation involved a quadraphonic speaker layout, so the top left of the image was mapped to the front left speaker, the bottom left to the rear left and likewise for the right side. When more than one region is present, the

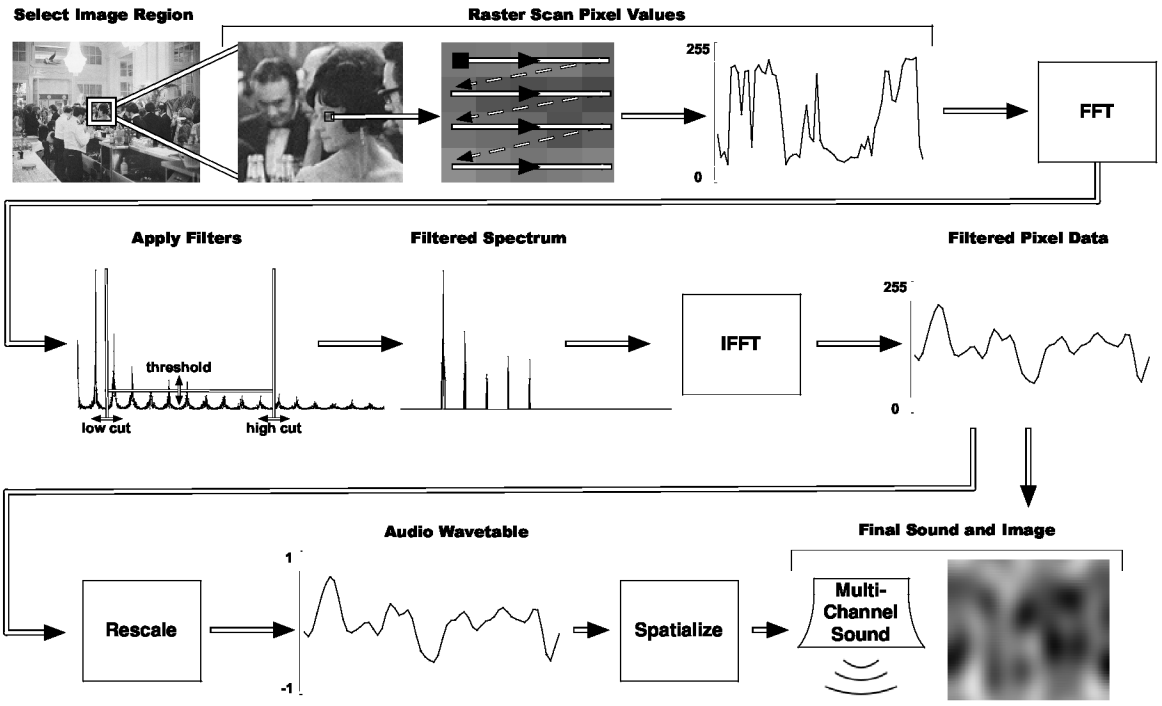

Figure 2: Sound Synthesis Algorithm for *Voice of Sisyphus*

spatialization provides a useful cue as to which sounds are coming from which regions.

The software is completely polyphonic– that is, one is free to add as many regions as desired, limited only by processor performance. Filtering and segmentation can be controlled independently for each region. Figure 4 shows an screen shot from *Voice of Sisyphus* containing 2 overlapping segmented regions producing different melodies. The screen shot also shows the software's ability to change opacity of the source image to hide or reveal it in the background.

Figure 4: Excerpt from *Voice of Sisyphus* Containing 2 Segmented Regions

2.4. Implementation

The software was programmed in C++ using OpenFrameworks[4], FFTW[5], and custom sound synthesis routines. OSC[6] allows for control of the software via any computer on the same network. In our case, the composer created a Java application with Processing that controls events in the composition. Using OSC it is also possible for many individuals to send control events to a single instance of the software, allowing a collaborative performance in which each individual controls his or her own region(s) of an image.

3. COMPOSITION

Voice of Sisyphus was conceived as an installation within a small gallery exhibition. Therefore, we wanted to compose it such that it would be compelling over a wide range of timescales. Specifically, it should have enough variation to reward prolonged engagement while also being diverse enough over the course of a couple minutes to give passing viewers a full experience.

The piece itself is an audiovisual composition that continuously cycles through a series of 8 phrases, each of which is used to convey a characteristic affect or mood. In order to prevent this repetition from becoming monotonous, the individual OSC commands or "notes" that sequence the events within each of these stages are not pre-defined but are generated in real-time based on a set of constraints. This choice ensured a theoretically limitless

[4] http://www.openframeworks.cc/
[5] http://www.fftw.org/
[6] http://opensoundcontrol.org/

number of variations over the same basic form throughout the duration that the piece was running.

Although our software supports the use of many simultaneously sonified regions, for the sake of simplicity we chose to limit ourselves to two, each of which play a distinct musical and visual role. The first region generally covers the entire width and height of our image and provides a stationary background over which moves the contrasting second region which selects smaller areas of focus. For the purposes of distinction we refer to them respectively as the "large" and "small" region.

3.1. Notation and Control

The Processing sketch that functions as our compositional score contains a series of parameters for each phrase that together determine its characteristic mood. Some values are set explicitly and remain constant though each cycle of the piece while others are chosen from within suggested constraints dynamically at the beginning of each section:

- Tempo dictates the speed at which new targets would be chosen for each parameter of the sonified regions (see quality below), as well as how quickly each region would jump about the screen if movement parameter 3 or 4 were selected. These range from 20 to 600 BPM and were set explicitly.
- Phrase lengths were chosen at random within 0.5 to 1.2 times a suggested value. The smallest suggested value was 15 seconds while the largest was 60, meaning that actual lengths range between 7.5 and 72 seconds.
- Movement type, which is the most distinguishing visual parameter, determines the way that each region moves over the image. Four main types of movement were defined, as well subtypes within each of these categories:

    1. Stationary: this was the type usually assigned to the large background region, which remains still while the small foreground region moves over it. During some sections both regions remain stationary.
    2. Smooth scanning: the region scans over the image either horizontally or vertically and either smoothly or in a randomized back-and-forth manner.
    3. Rectangular divisions: cycling randomly or in a sequential patterns in various directions, a region jumps over grid divisions of the image based upon powers of 2.
    4. Regions of interest selection: coordinates were manually gathered for all of the faces, groups of people, windows, glasses, lines, etc. within the image. These could be cycled through in various sequences.

For the regions of interest selection a "region group" variable controlled what type of object would be highlighted during each phrase or sub-phrase. For instance if *face* was selected, the smaller region would hop (on tempo) between ten pre-specified regions of the photo containing a person's face. Groups of people, windows, and glasses could each be highlighted in the same way, creating a total of 45 selectable features. The *line* setting, selects vertical strips of the image, corresponding to logical subdivisions of shapes within the scene. The ability to group visual information in this way demonstrates intent in what might otherwise appear to be a random system. Although these features are distorted and often difficult to identify, repetition suggests to the viewer semantical patterns, analogous to similar techniques used in film montage.

The fundamental frequency of both regions is set by tuning the scan rate in relation to the regions size. This tuning system provided a shorthand version of vertical harmony and values were chosen as simple ratios between the large background and small foreground region (1:2, 5:8, 10:1 etc.). Since the large region almost always remains stationary within the image and therefore has a fixed selection of pixels from which to derive frequencies, it functions like a slowly shifting drone or fixed bass over which the smaller region plays counterpoint as it moves through areas of different frequency content within the image.

Quality, which can be thought of similar to musical timbre, is actually a group of low-level filtering parameters that together determine a particular look and sound. As previously described, these include volume, mask, high-pass filter, low-pass filter, noise, and threshold. Each quality corresponds to series of suggested ranges for each of these parameters. At the beginning of a phrase, a smaller range of acceptable values is chosen from within this larger suggested range determined by the regions selected quality. While the phrase is playing, on each beat a new target is chosen from within this range of acceptable values, toward which the region interpolates. The speed at which this interpolation occurs is dictated by an independent parameter provided for the section. Large ranges of suggested values for each filter parameter are used to create phrases with a high range of timbres and quickly shifting forms, whereas a smaller range ensures that sights and sounds remain somewhat static. As a final method to ensure variation, some parameters are built in sets of 6 or 16 instead of 8, so exact repetition only occurs every 24 phrases, or 3 times through the cycle.

3.2. Compositional Themes

We wanted to depict abstracted and time-stretched methods of human/computer visual analysis. Specifically, we were interested in how an image is divided both geometrically and contextually, as well as how objects move between incoherence and recognizability. Each time our piece begins a new cycle, the entire scene is shown as a blurry and somewhat static mass of shapes. After some time the smaller region breaks off and begins to scan the image both smoothly and in jumping grid divisions. During this process, filtering of the underlying image continues to change, occasionally allowing viewers to distinguish faces and objects while at other times obscuring them completely. This is meant to mirror the way our eyes might initially try to make sense of a complicated scene. What are at first just masses of lines and shapes coalesce into identifiable forms. Likewise, as the smaller region eventually starts to directly target regions of interest within the image- faces, glasses, windows, lights, etc.- it mimics how we might scan different objects, categorizing them and placing them into logical groups based upon distinguishing features. At the climax of the piece, the entire image is shown unfiltered while the smaller region bounces as quickly as possible between important features in the image. After a few seconds this clarity fades back into a blur and the cycle starts once again.

3.3. Composition Through a Linked Audio-Visual Method

The compositional process was complicated immensely by the nature of our synthesis technique. Because sound material is generated directly by the pixels that comprise each sonified region, we

found that very interesting visuals often produced harsh or inappropriate sounds. Likewise, beautiful sonic harmonies could require very subdued or otherwise monotonous visual activity. For each section we were forced to run through countless variations and experiments in order to find parameters that produced unified and appropriate material in each sensory domain. However, as compensation for this difficulty, when the right settings are found, this method produces an audio-visual relationship that is perfectly synchronized and intuitively understood by unfamiliar audiences.

From a signal processing point of view, the results of the previous paragraph were not surprising. Non-acoustical data is inherently noisy when audified since it is not a time series of pressure data obeying the wave equation. Recognizable and meaningful visuals such as human faces are a complex arrangement of pixel values containing many frequencies when audified. The use of a variable amplitude threshold applied to the spectra of regions (outlined in section 2.3) allowed us to reduce noisy content of regions to obtain clearer, tonal sounds from otherwise complex regions. On the other hand, regions containing a simple arrangement of pixel values such architectural features (windows, edges of walls, lights, etc.), while less meaningful, lend themselves more naturally to coherent audification without heavy spectral modification.

In composing, we were not trying to substitute the visual modality of the image with a new sonic identity, but rather "add value" to the image in terms of Chion's "audio-visual contract," which describes how in film "we do not see the same thing when we also hear, and we do not hear the same thing when we also see."[12] The visual composition process of temporalizing a single image resulted in a sonic composition that in turn influenced modifications to our visual composition. The sound complemented and influenced the perception of the photograph to create an entirely new work of art. While clear portions of the image may produce otherwise unrelated abstract sounds (and vice versa), the audio-visual relationships are effective because of their precise synchronization and synthesis from the same data. We believe Chion's term "synchresis"[12] to describe a combination of *synchronism* and *synthesis* in film can also be used to describe our work.

4. INSTALLATION

Voice of Sisyphus was displayed at the Edward Cella Art+Architecture gallery in Los Angeles from November 5th, 2011 until February 4th, 2012. A single Mac Mini drove the projection and 4 channel sound for continuous operation during the aforementioned timeframe. Visitors to the gallery were free to move around the sound field or sit in a central point to experience the spatialization of the piece. After a few minutes of observation the title of the piece becomes understood. One reviewer put it, "As the image continuously reconstitutes itself and dissolves into a blurry abstraction the repetitive nature of Sisyphus' plight resonates."[7]

Most spectators quickly understood from the synchronized movement of image regions and sounds that various parts of the image were producing the audio track. However, one visitor commented on how well the chosen music matched the animation without realizing that the "music" was being generated from the image in real-time. We took this as a great complement for our sonification. Though the acoustics of the gallery space were far from ideal, the spatialization of the work proved to be quite effective as well. Listeners standing near the entrance were drawn to the center of the room once they heard the rapid movement of sounds.

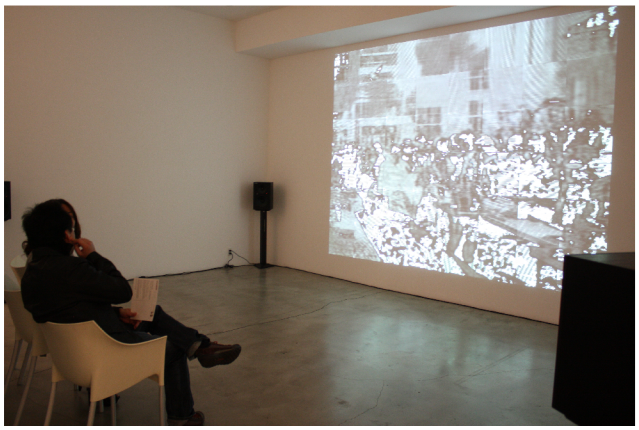

Figure 5: Installation of *Voice of Sisyphus*

5. FUTURE WORK

Future work in the area of visualization will be to implement automatic feature recognition to identify image regions of cultural interest (people, faces, etc.) using computer vision. The composition may then become autonomous from the sequential real-time sonification of image features as they are recognized by the computer. We are also interested the reverse– developing an algorithm that given a desired pitch or sound spectrum could find the best matching region within an image, thus automatically producing visuals for a composition.

6. REFERENCES

[1] B. Evans, "Foundations of a Visual Music," *Computer Music Journal*, vol. 29, no. 4, pp. 11–24, 2005.

[2] B. Schneider, "On Hearing Eyes and Seeing Ears: A Media Aesthetics of Relationships Between Sound and Image," *See this Sound: Audiovisuology 2*, pp. 174–199, 2011.

[3] C. Roads, "Graphic Sound Synthesis," *The Computer Music Tutorial*, pp. 329–334, 1996.

[4] W. S. Yeo and J. Berger, "A Framework for Designing Image Sonification Methods," in *Proceedings of International Conference on Auditory Display*, 2005.

[5] ——, "Raster Scanning : A New Approach to Image Sonification, Sound Visualization, Sound Analysis And Synthesis," in *Proceedings of the International Computer Music Conference*, 2006.

[6] ——, "Rasterpiece : a Cross-Modal Framework for Real-time Image Sonification, Sound Synthesis, and Multimedia Art," in *Proceedings of the International Computer Music Conference*, 2007.

[7] M. Wright and A. Freed, "Open Sound Control: A New Protocol for Communicating with Sound Synthesizers," in *Proceedings of International Computer Music Conference*, 1997, pp. 101–104.

[7] http://www.artillerymag.com/mini-reviews/entry.php?id=george-legrady-edward-cella-art-architecture

[8] A. Hunt and T. Hermann, "Interactive Sonification," *The Sonification Handbook*, pp. 273–296, 2011.

[9] T. Hermann, "Model-Based Sonification," *The Sonification Handbook*, pp. 399–425, 2011.

[10] B. Verplank, M. Mathews, and R. Shaw, "Scanned Synthesis," in *International Computer Music Conference*, 2000.

[11] W. Jones, "Sight for Sore Ears," *IEEE Spectrum*, February, 2004.

[12] M. Chion, W. Murch, and C. Gorbman, *Audio-Vision: Sound on Screen.* Columbia University Press, 1994.